IDENTIFIKACE AUTOMATICKÝCH PŘÍSTUPŮ INTERNETOVÝCH OBCHODŮ S VYUŽÍTÍM METOD WEB USAGE MININGU

Jana Filipová, Karel Michálek, Pavel Petr
Ústav systémového inženýrství a informatiky, Fakulta ekonomicko-správní,
Univerzita Pardubice

Abstract: *The paper presents models for identification of automatic access in the e-shop. Modeling of access is realized by neural networks and decision tree. The paper also deals with layout of architecture of web usage mining with Log files and on other side with JavaScript and cookies.*
Key words: Web mining, web usage mining, crawler, JavaScript, cookies

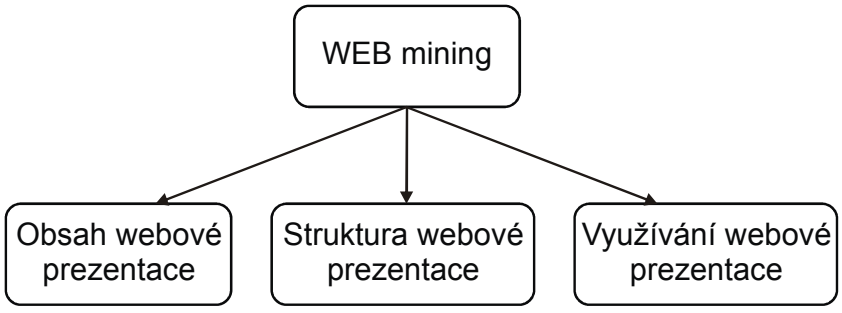
1 Úvod

V současné době velkou měrou narůstá obliba internetových obchodů (IO). V roce 2005 v České republice dosáhl obrat 75 největších internetových obchodů 10 miliard korun a tendence růstu je 50 % za rok [1]. Vzhledem k rostoucí konkurenci v této oblasti vyvstává potřeba internetových obchodníků účinně analyzovat chování svých zákazníků.

V IO je možné jednoduchými nástroji realizovat automatický sběr dat pro analýzu chování zákazníků. Tato data jsou ukládána do datových struktur umístěných v rámci nebo i mimo rámec IO. Nevýhodou automatického sběru dat v IO je však velké množství dat, která jsou v původní podobě pro jakékoliv manažerské závěry nevhodná. Proto vznikla nová aplikační oblast data miningu – web mining. Konkrétněji se jedná o web usage mining (dobývání znalostí na základě používání webu)[8].

2 Web usage mining

Cílem dobývání znalostí na základě používání webu je analýza chování uživatelů při využívání jednotlivých stránek [12]. Jedná se především o detekce vzorů v datech generovaných v průběhu spojení mezi klientem a serverem. V praxi se nejčastěji využívají statistické metody, asociační pravidla a metody segmentace (seskupování uživatelů s podobnými vzory chování nebo seskupování stránek navštěvovaných stejnou skupinou uživatelů). Zaiane a Han [8] uvádějí taxonomii web miningu, která je zobrazena níže (obr. 1).

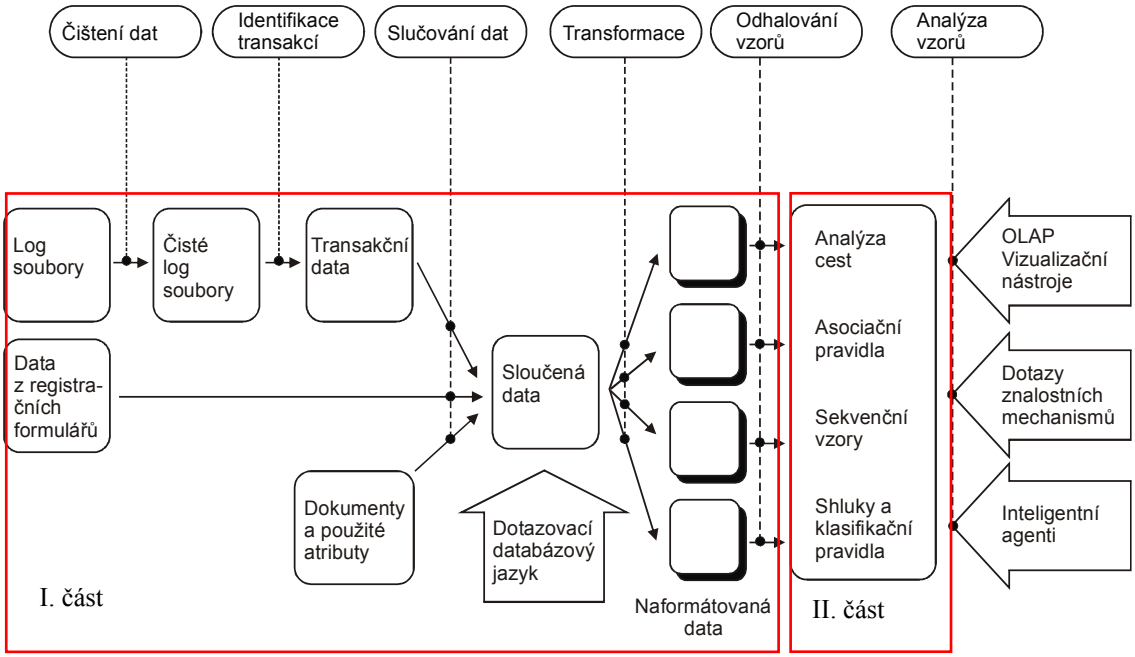


Obr.1: Taxonomie web miningu [8]

Data pro web usage miningový (WUM) rozbor lze získat z Log souborů nebo pomocí sofistikovanějšího způsobu a to kombinací cookies a JavaScriptu.

3 Architektura web usage miningu

Základní architektura WUM je publikovaná v [11] a [5]. Tato architektura rozděluje WUM na dvě základní části. První část zahrnuje transformaci webových dat, která se váží k doméně, do podoby využitelné pro zpracování (předzpracování transakčních identifikátorů a slučujících komponent). Druhá část zahrnuje jednotlivé metody data miningu (např. asociační pravidla a sekvenční vzory chování) jako součást systémového data miningu. Tuto architekturu web miningových procesů zachycuje obr. 2.



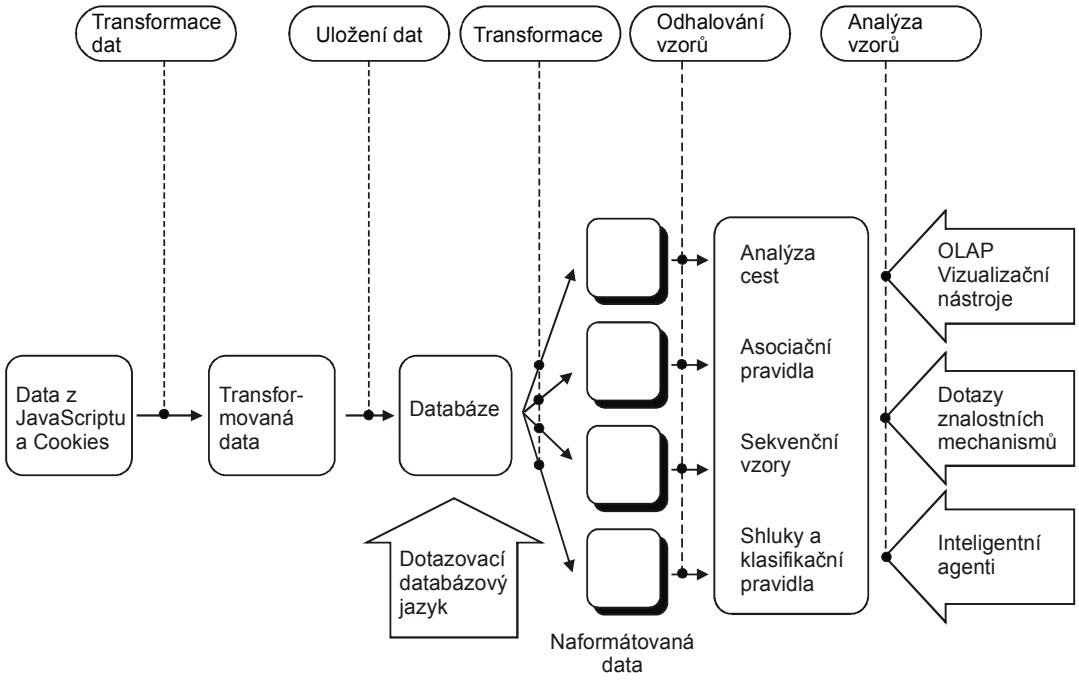
Obr. 2: Architektura Web usage miningu [6] se zdrojem dat z Log souboru

Log soubor obsahuje záznamy o všech požadavcích, které server zpracoval. Standardním formátem Log souboru je CLF (Common Log Format). V Log souboru se zaznamenává IP adresa, datum a čas návštěvy, frekvence návštěv a další informace [1].

Získávání dat pomocí JavaScriptu a cookies umožňuje získávat podrobnější informace o uživateli, než umožňují Log soubory a tím provést hlubší analýzu. Data získaná JavaScriptem jsou například informace o předchozí navštívené stránce, prohlížeči nebo o operačním systému uživatele apod. Z cookies je možné sledovat unikátní kroky uživatele v daném IO (zboží vložené do nákupního košíku, zaplacení u pokladny, atd.). Tyto údaje jsou agregovány do databáze na straně serveru. Nevýhodou tohoto řešení je velký počet operací s databází. Technologie sběru dat pomocí JavaScriptu a cookies uložených do databáze mění předchozí model (obr. 2) následujícím způsobem (obr. 3).

Z obr. 3 vyplývá, že tento model je jednodušší pro vlastní data miningovou analýzu tím, že je oproštěn o komplikovanou transformaci webových dat a jejich následné formátování (stačí pouze napojení do databáze například pomocí SQL).

Běžné parametry sledované u internetových obchodů jsou např.: návštěvnost, konverzní poměr přístupu, konverzní poměr objednávek, technická analýza na straně klienta nebo ROI internetových kampaní.



Obr. 3: Architektura Web usage miningu [6]se zdrojem dat z JavaScriptu a cookies

4 Automatické přístupy a crawler

Web usage mining pomocí jeho nástrojů umožňuje daleko hlubší zkoumání chování uživatelů, než jsou jen běžné operace s databází. Web usage mining realizuje v první fázi problematiku čištění získaných dat od zavádějících návštěv internetových vyhledávačů tzv. automatických přístupů. Automatické přístupy vytvářejí tzv. web crawlers (web robots, web spiders). Jedná se o druh programu, který automaticky prochází www prostorem a jeho hlavním účelem je vytvářet kopii navštívených webových stránek pro internetové vyhledávače, které jsou indexovány v jejich databázích. Záznamy o pohybu crawleru na webové stránce jsou v databázi zaznamenány stejně a téměř nerozeznatelně od pohybu konvenčního uživatele.

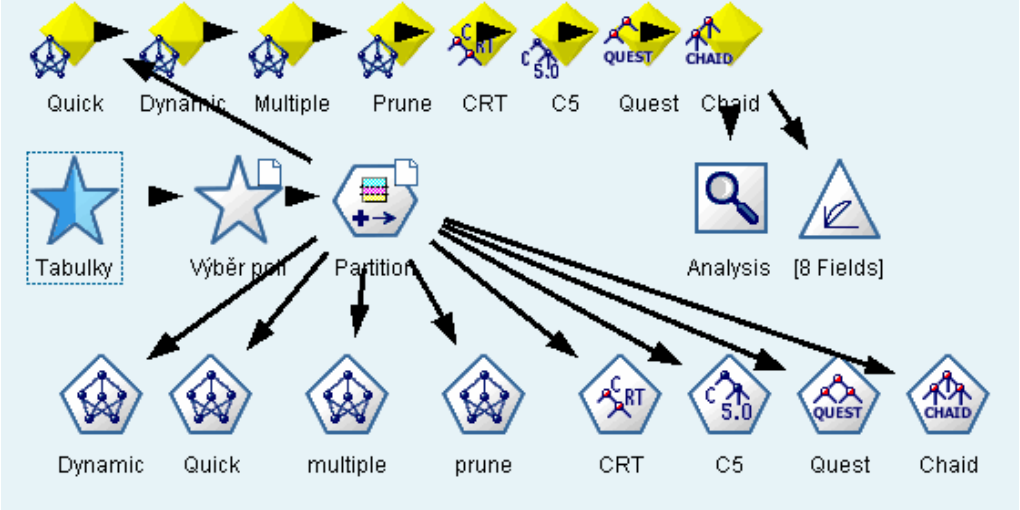
Automatický přístup nelze v databázích určených pro WUM jednoduše identifikovat. Jeden z možných způsobů identifikace automatického přístupu je možné provést pomocí IP adresy crawleru. Tento způsob vyhledávání není zcela přesný, protože crawlerů je celá řada, stále vznikají nové (např. Wikipedie se zmiňuje o 25 různých crawlerech) a jejich počet neustále narůstá s rozšiřujícím se trhem s fulltextovým vyhledáváním [13]. Crawlery používají celá pásma IP adres (google crawler využívá řádově tisíce IP adres) nebo se jejich IP adresy mohou v čase měnit. Při identifikaci crawleru pomocí IP adresy je nevyhnutné neustále udržovat aktuální databázi crawlerů. To však vzhledem k výše popsanému není možné.

Další možností jak automatický přístup identifikovat je podle toho, zda daný přístup měl založeno cookies (crawlery cookies nemohou zakládat). Dalším parametrem pro identifikaci je, že crawler má větší počet načtených stránek v čase než konvenční návštěvník. Hranice může byt stanovena mezi 10 – 20 načtení stránek za minutu. Tento způsob identifikace crawlerů může být ale v některých případech zavádějící. Mohou existovat konvenční návštěvníci, kteří mají zakázané cookies v prohlížeči, např. z bezpečnostních důvodu. Dále mohou provést vyšší počet obnovení stránek, např. díky nekvalitní konektivitě internetového připojení. Tito návštěvníci potom mohou být označeni jako crawler a nemohou na daném IO často dokončit svůj nákup. Právě zde se nabízí prostor pro využití data miningových metod pro odhalování automatických přístupů.

5 Modely pro detekci automatických přístupů

Pro návrh a analýzu modelu detekce automatických přístupů byla použita data konkrétního internetového obchodního domu. Doposud firma identifikovala automatický přístup podle toho, že tento přístup neměl cookies, tudíž nemohl být ani evidován v tabulce logů. Druhým parametrem pro identifikaci bylo, že automatický přístup má větší počet obnovení stránek (reload) než fyzický návštěvník. Hranice byla stanovena na 10 obnovení. Tato hranice se však v praxi ukázala jako velmi nízká hranice, protože spolu s automatickými přístupy byli označeni i fyzičtí návštěvníci. V tom případě byl potom těmto návštěvníkům znemožněn nákup v IO. Podmínkou pro dokončení objednávky a nákup v tomto IO je, že každý zákazník musí mít povolena cookies.

Pro tvorbu modelů automatické identifikace automatických přístupů bylo využito následujících tabulek: tabulka Log soborů, tabulka IP adres, tabulka prohlížečů, tabulka jednotlivých přístupů (sessions) a tabulka detailů. Proces zpracování a návrhu modelů je znázorněn na obr. 4. Uzel datového proudu označený jako Tabulky realizuje propojení jednotlivých tabulek z databáze IO. V následujícím uzlu Výběr polí jsou selekována jednotlivá pole a určen jejich příznak (vstupní, výstupní pole) (obr. 5). Dalším uzlem je uzel Partition, který dělí vstupní množinu dat na trénovací, testovací a ověřovací množinu v poměru 40:30:30. Na tento uzel navazují uzly tvorby modelů (dolní část obr. 4) a poté jednotlivé naučené modely (vrchní část obr. 4).



Obr. 4: Datový proud

Field	Type	Values	Missing	Check	Direction
Record_Count	Range	[1,454]		None	In
is_no_cookie	Flag	1/0		None	In
count_use	Range	[1,6486]		None	In
agent_group	Set	"999999",...		None	In
op_sys	Set	"999999",...		None	In
agent_version	Set	"999999",...		None	In
bot	Flag	1/0		None	Out

Obr. 5: Výstup z uzlu

Výběr polí

Pro odhalování automatických přístupů byly zvoleny dva typy modelů - rozhodovací stromy a neuronové sítě. Vstupní data, která byla použita pro další analýzu jsou (obr.5): počet přístupů (Rekord_Count), cookies (zda-li má uživatel povoleno cookies nebo ne), počet obnovení stránky – aktualizací (count use), typ prohlížeče, typ operačního systému, přesný název prohlížeče a jeho verze. Výstupním polem je příznak, zda daný přístup je automatický (1) či konvenční návštěvník (0), toto pole je označováno jako bot.

5.1 Modely neuronové sítě

V případě návrhu neuronové sítě (NS) byly využity čtyři trénovací metody pro návrh konfigurace NS (Quick, Dynamic, Multiple a Prune). Topologii NS určuje metoda návrhu NS a struktura trénovacích dat. Vstupní vrstva neuronů je tvořena vstupními atributy a výstupní vrstva je tvořena jedním neuronem reprezentujícím závěr, zda je daný přístup automatický nebo ne (1/0).

Metoda Quick využívá „quick“ metodu, která používá pravidla a charakteristiky dat pro vybrání nejvhodnější topologie.

Metoda Multiple vytváří více sítí různých topologií (přesné číslo závisí na trénovacích datech). Po natrénování je vybrán model s nejmenší chybou.

Metoda Prune využívá metodu postupného odřezávání nejslabších jednotek ve vstupní a skryté vrstvě NS.

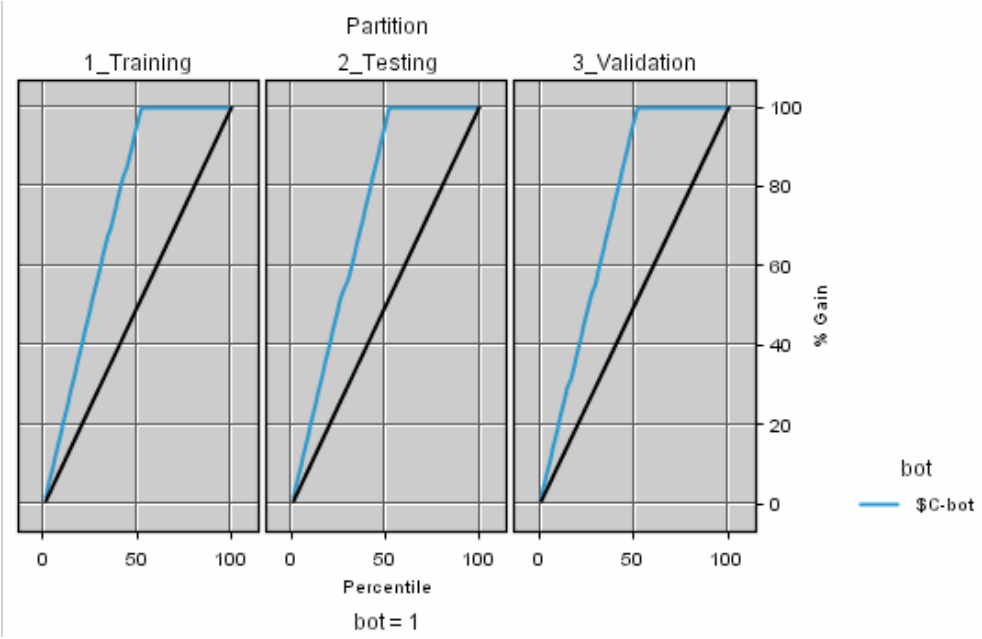
Metoda Dynamic modifikuje standardní strukturu NS přidáním nebo odstraněním uzlů v skryté vrstvě v průběhu procesu učení.

Použití neuronových sítí s sebou však přináší problém v interpretaci výsledků. Z neuronové sítě se stává „černá skříňka“, do které laik „nevidí“. Objevují se ale pokusy převést znalosti uložené v neuronových sítích do srozumitelnější podoby.

5.2 Model rozhodovacího stromu

Pro tvorbu rozhodovacího stromu byl vybrán algoritmus CART [3, 4] (v každém nelistovém uzlu se data rozdělují do dvou množin), algoritmus C5 [3, 9](umožňuje práci s numerickými atributy, chybějícími hodnotami, převod na pravidla i prořezávání) a algoritmy Quest [10] a Chaid [3, 4]. Výhodou rozhodovacího stromu je, že jeho řešení lze převést na rozhodovací pravidla.

Pro volbu nejvhodnější metody bylo provedeno srovnání pomocí grafu Evaluation (obr. 6). Tento graf nabízí jednoduchou možnost porovnání modelů pro výběr nejvhodnějšího z nich (porovnává náhodný výběr s výběrem zvoleného modelu a dané vstupní množiny dat).



Obr. 6: Graf trénování, testování a ověřování rozhodovacího stromu C5

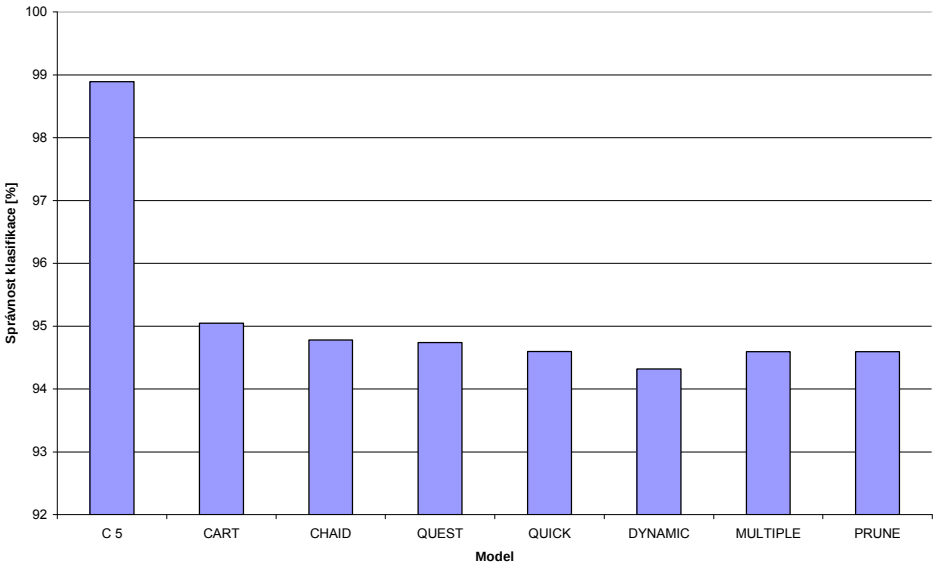
6 Závěr

Na základě realizovaných analýz (porovnání výsledků NS a rozhodovacích stromů) vyplynuly následující závěry (Tab. 1).

Tab. 1: Výsledky klasifikace jednotlivých modelů

Model	Správnost klasifikace [%]	Model	Správnost klasifikace [%]
C 5	98,89	QUICK	94,6
CART	95,05	DYNAMIC	94,32
CHAID	94,78	MULTIPLE	94,59
QUEST	94,74	PRUNE	94,59

Jednotlivé metody dávají srovnatelné výsledky (obr. 7), (tab. 1). Nejlepšího výsledku ale dosáhla metoda rozhodovacího stromu C 5 (obr. 8).



Obr. 7: Graf výsledků klasifikace automatických přístupů

Individual Models

Comparing $C-bot with bot

'Partition'	1_Training		2_Testing		3_Validation	
Correct	66652	98,87%	50086	98,81%	50239	98,89%
Wrong	763	1,13%	603	1,19%	566	1,11%
Total	67415		50689		50805	

Obr. 8: Dosažená úspěšnost modelu C5

Pro praktické použití je v tomto případě vhodnější využít model na základě rozhodovacího stromu C 5. Hlavním přínosem použití rozhodovacího stromu je jeho jednoduchá interpretace ve formě pravidel (obr. 9) [8]. Tato pravidla se potom dají implementovat do prostředí daného IO za účelem detekce automatických přístupů.

```
agent_group in ["999999" "Gecko" "MSIE" "Opera"] [Mode: 0] ⇨ 0
agent_group in ["Other"] [Mode: 1]
    count_use <= 1197.500 [Mode: 1]
        count_use <= 3.500 [Mode: 0] ⇨ 0
        count_use > 3.500 [Mode: 1] ⇨ 1
    count_use > 1197.500 [Mode: 0] ⇨ 0
```

Obr. 9: Ukázka rozhodovacích pravidel modelu CART

Jako DM prostředí byl zvolen software od společnosti SPSS Clementine 9.0, který podporuje metodologii DM-CRISP [8] podle které bylo také při analýze postupováno.

Literatura:

[1] BERKA, P. Dobývání znalostí z databází. Praha: Academia, 2003. s. 400, ISBN 80-200-1062-9.

[2] BERKA, P. Rozhodovací stromy: [online]. [cit. 2006-5-22]. Dostupný z WWW: <http://lisp.vse.cz/~berka/docs/SL-WEB.PDF >

[3] BERSON, A., SMITH, S. J. Data Warehousing, Data Mining & OLAP, McGraw-Hill, USA, 1997.

[4] BERRY, M. J. A., LINOFF, G. S. Data mining techniques: For Marketing, Sales, and Customer Relationship Management, Second Edition, Wiley Publishing, Indianapolis, Indiana, 2004.

[5] COOLEY, R., MOBASHER, B., SRIVASTAVA J. Grouping web page references into transactions for mining world wide web browsing patterns. Technical Report TR 97-021, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.

[6] COOLEY, R., MOBASHER, B., SRIVASTAVA, J. Web Mining: Information and Pattern Discovery on the World Wide Web. BAMSHAD MOBASHER [online]. 1997 [cit. 2006-10-30]. Dostupný z WWW: <http://maya.cs.depaul.edu/~mobasher/webminer/survey/survey.html>.

[7] E-commerce v ČR - První pohled zblízka. APEK [online]. 2006, 25.5.2006 - GfK Praha [cit. 2006-10-30]. Dostupný z WWW: <http://www.apek.cz/tiskove-informace/tiskove-zpravy/e-commerce-v-cr-prvni-pohled-zblizka/>.

[8] CHAPMAN, P., et. al. CRISP- DM 1.0 Step-by-step data mining guide. SPSS inc. 2000. [on-line]. [cit. 2006-5-05]. Dostupné z URL http://www.crisp-dm.org/

[9] MACHOVÁ, K. Strojové učenie: Princípy a algoritmy, [online] [cit. 2006-04-10], Dostupný z WWW: <http://neuron-ai.tuke.sk/~machova/>, Technická univerzita v Košiciach.

[10] Manuál programu SPSS 14.0 for Windows, [online] [cit. 2006-04-29], Dostupný z WWW: <http://www.spss.com/spss>.

[11] MOBASHER, B., JAIN, N., HAN, E., SRIVASTAVA J. Web mining: Pattern discovery from world wide web transactions. Technical Report TR 96-050, University of Minnesota, Dept. of Computer Science, Minneapolis, 1996.

[12] SRIVASTAVA, J., COOLEY, R., RESHPANDE M., TAN P. Web usage mining: discovery and applications of web usage patterns from web data. SIGKDD Explorations, Vol.1, Issue 1, 2000.

[13] Web crawler: Examples of Web crawlers. [online]. 2006 [cit. 2006-10-30]. Dostupný z WWW: <http://en.wikipedia.org/wiki/Web_crawler>.

[14] ZAIANE O., HAN, J. WebML: Querzing the World_wide Web for resources and knowledge. In: Proc. Int. Workshop on Web Information and Data Management WIDM´98, Bethesda, 1998.

Kontaktní adresy:

Ing. Jana Filipová, Ing. Karel Michálek, DiS., doc. Ing. Pavel Petr, Ph.D.
Ústav systémového inženýrství a informatiky
Fakulta ekonomicko-správní
Univerzita Pardubice
Studentská 84, 532 10 Pardubice
jana.filip@gmail.com,michalek@informacni.org,pavel.petr@upce.cz